下载APP

11 | 因果倒置：星座真的可以判定你的性格吗？

2021-08-20 郭炜

《数据分析思维课》 课程介绍 >

讲述：郭炜

时长 18:55 大小 17.33M

数据给你一双看透本质的眼睛，这里是《数据分析思维课》，我是郭炜。

前面给你讲了很多数据分析的基本方法，在有了这些基本方法之后，其实还有很重要的一步是确定数据背后的逻辑，否则分析数据就和算命一样，看了下手相就直接告诉你一个结论。比如以下根据“数据分析”形成的耳熟能详的结论，你想想看究竟有多少是值得推敲的呢：

学术派：打篮球会让人长高，喝咖啡能长寿，不吃早饭会导致肥胖；

网红派：爱笑的女孩子通常运气都不会太差（很暖心，对吧），《奇葩说》讨论的“撒娇的女人更好命”。

这些结论感觉上都很有道理，甚至很多还有一些数据统计的报告来作为佐证，但其实仔细看背后的数据和逻辑，往往缺乏依据或者会出现因果倒置、因果无关的这种情况。

这种对于数据的用法其实是最危险的，因为这里面的问题往往隐藏得非常巧妙，你如果不深究其中的逻辑，往往会被数据所欺骗从而得出错误的结论，甚至指导你进行一些错误的行为。

所以我在这节课里，为你总结了最常见的 6 种误用数据导致错误因果结论的陷阱，希望你在今后的工作和学习当中，不要因为学了分析数据，最后却在分析数据时被数据所欺骗。

相关因果倒置——鸡叫导致天明

第一种就是最常见的因果倒置，这在公务员考试里也有专门的考题。

典型的因果倒置就是天亮了鸡就开始打鸣，但是我们不能说是因为鸡打鸣导致了天亮。但是实际的应用当中，我们往往会忽略这个逻辑。比如，我们在一些医学统计上会看到说不吃早饭会导致人肥胖，甚至还有大量的统计数据表明这些肥胖的人都没有吃早饭。

问题是，数据的确是同步发生的，但是不代表这些数据之间有因果关系。而且有可能会出现因果倒置——肥胖的人胖所以早上不饿，所以他不吃早饭。而比较瘦的人自身代谢比较快，晚上消耗多，早上就会比较饿，所以他要吃早饭。

所以如果你没有了解这个原因，然后只是很简单地觉得吃早饭就不会变胖（然后早上包子油条糖油饼），可想而知你的体重肯定涨的更快了。

同样，很多统计数据表明，选择多种投资方式的人往往比只从事本职工作的人更富有，现在都流行做投资，做一个有“睡后”收入的人。

相应地，很多人会认为多种投资方式会让人更加富有，其实这也是因果倒置。实际上只有人们拥有一定财富之后，才会去选择多种方式投资。所以如果你现在很穷，你直接去尝试多种投资方式，你大概率不会更加富有，有可能赔得更多了。

往往由于我们对事实的逻辑不清楚，我们会把事件的结果当成原因，这就会导致我们得到一些荒唐的结论（鸡打鸣导致天亮）。最终，如果我们按照这个数据结论进行操作，往往得不到我们想要的结果，还可能造成严重的危害。

所以当我们看到数据结果的时候，一定要仔细推敲其中的业务逻辑，同时进行反向测试。例如对于吃早饭是不是会让人减肥这件事，我们就可以选两组类似的人去做随机对照实验，一拨人吃早饭，一拨人不吃早饭，其他的时间用餐的量和他们活动的量都一致，最终看不吃早饭到底能不能够减肥。如果发现他们的体重没有太大差异，那就说明吃不吃早餐和减肥之间没有任何因果关系。

相关性而非因果——吸烟真的致癌么

第二种我们经常看到的数据分析结论错误是数据相关而并非因果。这里面的例子非常多，比如曾经流行一个说法叫做喝咖啡能够长寿，海外和国内的官方报道都报道过此事。

人民网 people.cn 人民网 >> 四川频道

16年间70多万人研究表明：每天喝咖啡能长寿！

2017年07月20日10:31 来源：人民网-健康时报 分享到：

原标题：16年间70多万人研究表明：每天喝咖啡能长寿！

但结论未必如此。因果是一个非常强的逻辑，我们初中学过因果叫充分条件，而不是必要条件。也就是说，因果意味着我们如果做了 A，那么一定会导致 B 的发生。这在数据的领域里面其实是非常难证明的，我们可以通过数据实验去证明 B 发生和 A 没有关系，但是很难证明 A 就是 B 的发生的充分条件也就是原因，因为有可能他们之间只是数据相关关系，而不是因果关系。

是不是感觉有点晕？没关系，现在我用一个你非常熟悉的例子，再给你解释一下。我们经常能看到吸烟会致癌这个理论。但是吸烟真的能致癌吗？

当然，我不是要替吸烟的人辩护，从健康程度来讲，吸烟的确是有害于健康的。但是从科学角度上来讲，尽管医学家、统计学家在过去的几十年里做了非常多的试验，但到目前为止，我们还没有确凿的统计学证据可以说明吸烟致癌。因为致癌的因素太多了，你无法判断吸烟能够直接导致癌症。

现代统计学的奠基人费舍尔对香烟会导致肺癌结论表示了强烈的质疑，他只确认了吸烟和患有肺癌之间有相关性，但是从科学的角度上来讲，的确不能说因为吸烟，所以会导致肺癌。

看上去数据是正确的，但是如何解释数据其实非常需要动脑子。两件事情虽然相关，但是往往无法说明它们之间有因果关系。而因为我们的大脑容易记住有逻辑性的东西，所以我们经常把相关的东西“套上”一个因果的外壳，但这其实是不对的。

因果关系需要大量实验的验证证明，只要有 A，B 一定会发生（且没有其他因素的干扰）才能说明 A 可以导致 B 的发生。而这在现实里其实是非常难做到的，就像前面提到的吸烟导致肺癌，如果要严格意义证明这一点，我们必须要找到若干组同卵双胞胎（确保他们的基因类似），饮食结构完全一样，活动也完全一样。然后让他们一组人吸烟，另外一组不吸烟，还要让他们相互不知道在测试（确保对照试验公正性），最后还得吸烟组得了肺癌才有可能通过数据证明吸烟真的会导致肺癌。

这个困难度可想而知。**所以以后在工作和生活中，不要轻易下因果关系的结论，相关并非因果。**

遗漏 X 变量——找到背后真实原因

我们在做数据探查之后发现了几个数据之间的相关性，虽然我们无法确认它们是因果关系，但真正在数据的探查和分析过程中，我们很有可能会找到相关的真实原因，从而去解决问题。这里我把它叫做找到遗漏的 X 变量。

举一个很有意思的例子，在英国瓦努阿图岛上的居民有一个奇怪的信仰：他们相信虱子有益于身体健康。因为经过数百年的观察，这里人的发现身体健康的人身上通常都有虱子，而生病的人就没有。

数据本身是准确无误的，科学家们也发现同样的数据，但不代表着岛上居民这个“虱子让人健康”的信仰就正确。后来经过自然学家实地考察发现了真相，原来这里几乎所有人的身上都有虱子，但是如果有人发烧，随着体温升高，虱子会因为受不了高体温而离开。

所以看上去是虱子使人健康，其实是体温高导致虱子不栖息在人身上。所以，岛上居民的结论应该是看到没有虱子的人应该让他去就医，因为他发生了疾病。在原始部落并没有体温计，这个结论的确可以帮助到他们，而不是盲目的相信“虱子让人健康”。

再给你举个例子，现在都非常讲究母乳喂养，但究竟母乳喂养应该多久呢？世界卫生组织在《婴幼儿喂养指南》中建议母乳喂养两年或更长的时间。相关研究也表明，与非母乳喂养的婴幼儿相比，母乳喂养的婴幼儿患某些传染病的风险更低，也有更低的死亡率。

然而，在一些研究中研究人员发现，对接受母乳喂养时间更长的婴幼儿来说，营养不良的风险更高。这是对的么？应该缩短母乳喂养的时间么？1997 年，来自美国约翰斯·霍普金斯大学的研究人员专门就此进行了分析，发现真实原因是收入比较低的家庭通常其他食物非常有限，更倾向于接受更长时间的单一母乳喂养。

所以没有充足的辅食才会导致婴儿营养不良，现在新生儿家庭都已经知道除了母乳喂养，在后期要增加辅食才可以让孩子更健康。

所以，**当我们在日常生活和工作当中看到两个数据强相关的时候，即使不能把它们当成因果关系，也可以顺藤摸瓜找到可能的原因，再用业务逻辑或者实验去验证这个可能的原因是否为真实原因。**缺乏业务逻辑的数据，永远只会是数据。缺乏数据的业务逻辑，也永远只是在纸上的一个业务逻辑图而已。

即使我们在找对了两件事情前后的因果关系，我们也会因为整体的选择对象、覆盖范围以及时间长度导致出现因果关系的推断错误。下面我再分别给你讲解一下这三种误区。

以偏概全——伯克松悖论

第一个就是统计数据本身因果逻辑成立，但是以偏概全。统计学里对这个现象有一个特别著名的理论叫做“伯克松悖论”。

伯克松悖论指的是当不同个体被纳入研究样本的机会不同时，研究样本中的两个变量 X 和 Y 表现出统计相关，而总体中 X 和 Y 却不存在这种相关性。听上去是不是有点拗口？没关系，我举两个具体例子来帮助你理解。

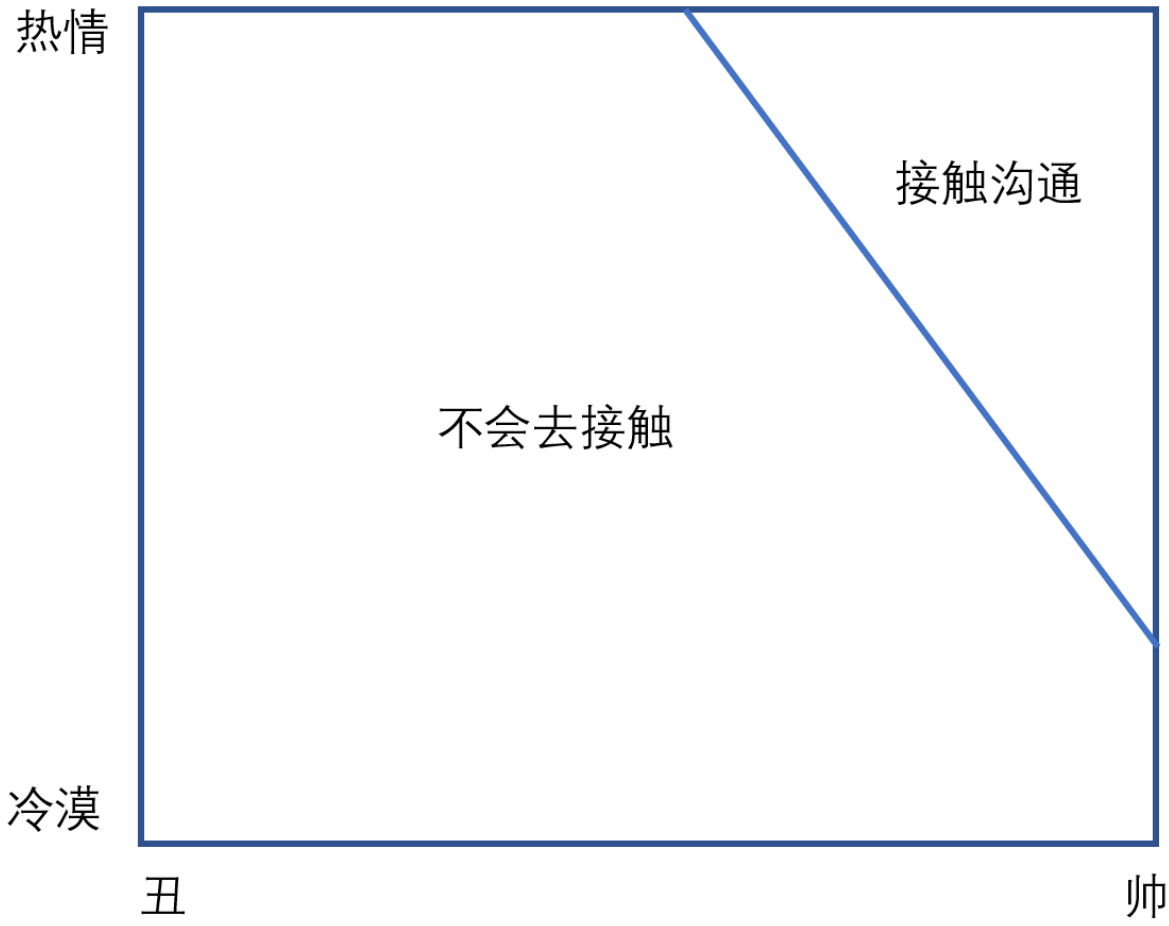
第一个例子是著名的“海军与平民死亡率”的例子。在 1898 年“美西战争” 期间，美国海军的死亡率是 9%，而同期纽约市市民的死亡率为 16%。后来海军征兵部门就拿这个数据跟大家讲，待在部队里其实比大家待在家中更加安全。

这逻辑肯定是错误的，但是错误不在具体数据，而是这两组数据其实没有什么可比性。因为海军的主要是年轻人，他们身强体壮、不会出现太多身体疾病；而纽约市民里面包含了

新出生的婴儿、老年人、病人等等，这些人无论放在哪里，他的死亡率都会高于普通人。

所以，参军不能说比大家待在家中更加安全，但反过来你也无法证明待在家中就比参军更安全，因为比对的对象不是在同一个人群里，这就是伯克森悖论。

同样，现在的城市女孩会觉得，对他很热情的男生往往都长得不帅，长得很帅的男生往往对她都不够热情。但其实帅和不帅并不是导致男孩热不热情的原因，只是因为只有长得帅或者对女生够热情的男生，才有更多机会和女孩子接触，你看下图就明白了，还是从局部看整体的逻辑不对。



在我们工作当中也会遇到这样的情况。我们经常会通过一些调查问卷去访谈一些使用者，看看营销效果如何。

比如，现在我们有一个用户访谈的数据结果：购买某品牌产品的 100 人中，有 72% 的人说在一个月内看过该品牌的广告 ；而在未购买商品的 300 人中，有 76% 的人说一个月内没看到过这个品牌的广告。

通过这个数据我们能看到什么呢？我们可以下结论说是广告提高了我们的用户转化率吗？这是不可以的，因为实际购买的人会对广告更有印象，而没有购买的人也许也看了广告，只不过他没有印象而已。因为统计范围不同，所以不能够根据这个数据给出转化率比较高的结论，然后大肆提高广告投放。

看上去有因果关系的数据，还要看数据集的比对性才可以给出数据最终的结果。

控制数据范围——神枪手谬误

前面给你讲散点图的时候，讲过神枪手谬误，这是一个典型的控制数据范围导致错误的数据结论逻辑。我们在生活中也很容易遇到这种陷阱，所以我在这里再给你强调一下。

很多统计结果其实是被操纵的，他们把某些机缘巧合之下比较好的结果的相关数据放到一起，去证明一个不可能的事情，但是如果你再换一组数据，那么你就没有办法证明这个因果关系。例如曾经在国内炒的火爆的全国牙防组故事就和这样的数据有关。

在海外也有一些小众的牙膏制造商，为了证明自己的牙膏比其他牙膏有效果，只把好的结果公之于众。包括很多“伪学术论文”引用的数据，也不是多次统计的结果，而是选取最优的结论给出来。

所以你在看最终数据分析报告的时候，一定要看它的数据是不是先有枪眼再画靶子，或者先找到满意的结果再给你看统计数据，我们需要的是通过大量的随机样本给出的结果。

时间长度不足——替代终点问题

还有一些数据在分析和统计的时候，由于时间长度不够，会造成数据统计的结果不准确。这个在学术上我们叫做“替代终点问题”（ surrogate endpoint problem ）。

比如我们要检测某种药物是不是可以延年益寿，这其实就需要投入大量的时间和资金，因为我们必须得等到人们去世以后才能知道他们的寿命。

所以对于现在各种各样的保健品，如果它的宣传的作用是可以延年益寿的话，那大部分都是收你“智商税”的，因为这种测试非常难以完全实现。即使服用这些药物的人最后长寿了，那也不能够代表这两种之间存在着因果关系，很有可能只是前面讲到的相关性。

同理，你看这么多风险投资人在选择创业公司的时候，其实是**靠大的方向和辨识团队来进行投资，而不是靠具体某些数据来表明这个创业公司是否靠谱。**因为相对一个创业公司来讲，公司的成立时间太短了，公司的数据不代表趋势，这就是替代终点问题。

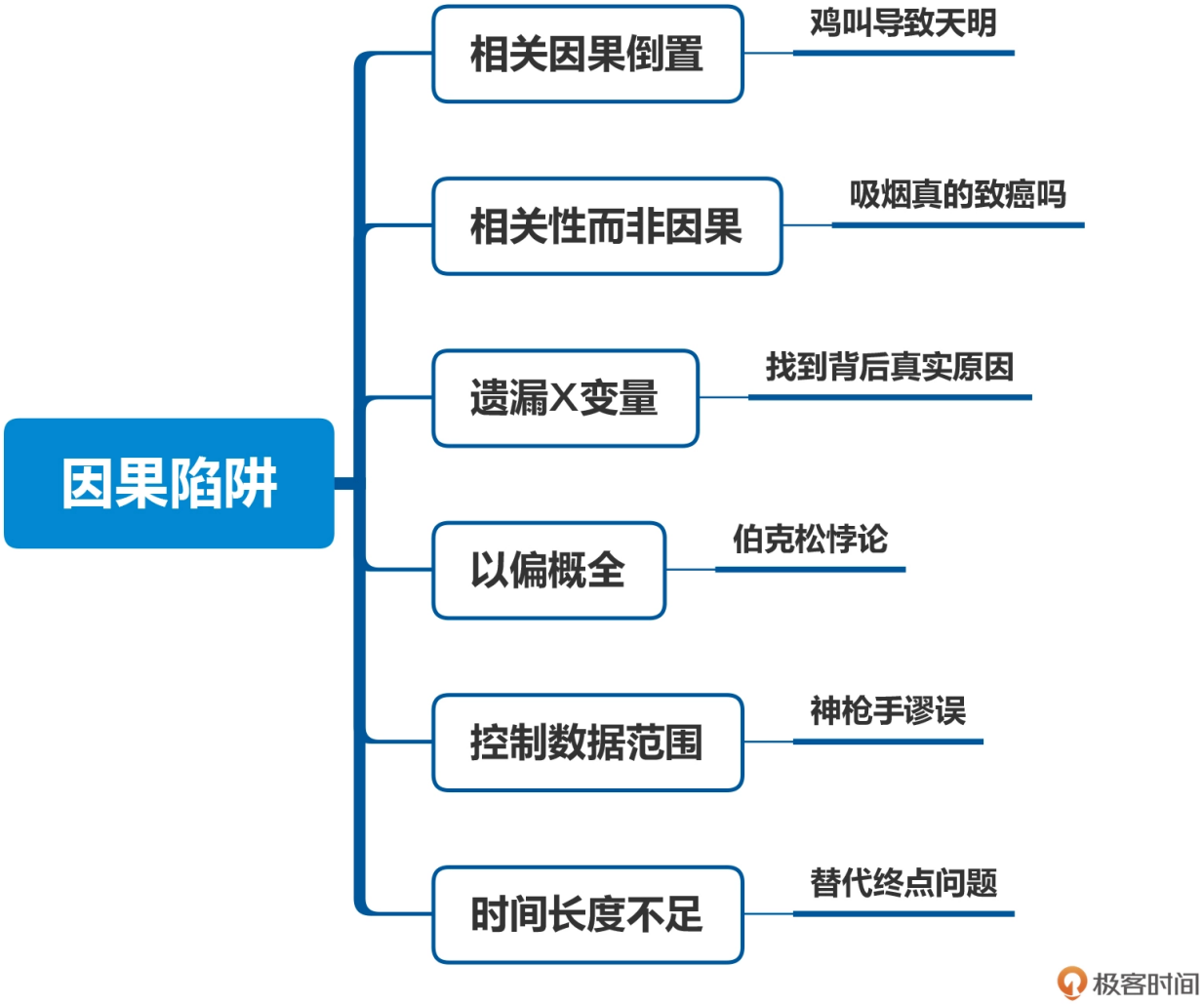
小结

学了这些场景以后，我们回过头来再看看开头的那些问题。

> 打篮球真的能让人长高吗？这很有可能是因为长高的人都会去打篮球，而不是打篮球让人长高——因果倒置。
>
> 喝咖啡可以长寿？常喝咖啡的人一般都是白领阶级，他们的营养供给更高，所以他们可以长寿，而不是因为咖啡让他们长寿——相关性而非因果关系。
>
> 吃不吃早饭其实和你肥不肥胖没有什么关系，运动健康才和你的肥胖有关系——相关性而非因果关系。
>
> 爱笑的女孩子通常运气都不会太差？爱笑的女孩其实运气也有差的，最后她就不笑了，事实是因为运气好的女孩她们才会爱笑——因果倒置。
>
> 会撒娇的女人更好命？女人好不好命其实与另一半或者周围的人和环境更有关系，而不是和你会不会撒娇有关系——需要找到遗漏的 X 变量。

这节课是我们数据分析基础篇的最后一节课了。在前面的课程里，我给你讲了非常多的数据统计的方法，你可以很快地把这些数据分析方法应用到自己的工作当中。今天我们其实是换了个思路，给你主要讲的是数据本身的局限性。数字相关并不等于因果关系，对于做数据分析和做数据决策来讲，我们更要懂业务才能够去了解真相，不然很容易就被数据忽悠了。

数据分析就像是一门中西医结合的医学，既要有本章前面给你的这些数据分析办法，也要有接下来的章节会讲到的算法模型和工具。最终还是需要你这个人像老中医一样，能够对这个业务本质有深刻的理解和把握，才能给出最终正确的结论。让我们一起持续学习，一起共勉。



课后思考

最后还是我们的课后思考环节。我们这节课里讲了这么多因果相关性的问题，最后回到今天的标题：星座真的可以判定你的性格吗？你觉得它是一个什么样的问题呢？欢迎你在评论区一起讨论，加入到“中西医结合”的数据分析行业里来。

分享给需要的人，Ta订阅后你可得 20 元现金奖励

赞 3 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 10 | 指数和KPI：智商是怎么计算出来的？

下一篇 12 | 精神鉴与看信仪准：两种预测，你会首先相信哪一个？

精选留言 (5)

写留言

六月天 良

2021-08-20

星座判断人的性格，以偏概全，和广告投放的例子类似。星座判断的比较准的人会告诉别人，不准的人一般不会宣传出来。得到的结果是星座判断人的性格很准。

另外，星座判断人的性格，有些性格写的比较模棱两可，让人看了之后会往上想，确实是那样。

展开

作者回复: 是的，算命的手法

5

Flychen 神

2021-08-23

缺乏业务逻辑的数据，永远只会是数据。
缺乏数据的业务逻辑，也永远只是在纸上的一个业务逻辑图而已。

2

魔

2021-08-20

这节课妥妥的逻辑课呀，受益良多！

星座决定性格，听莫名其妙的，甚至有些人还以星座来作为择偶的标准呢。星座很八卦，有点玄学的意思，神乎其神却好似怎么都说得通。稍微不动点脑筋，还真的信以为真了~

展开

魔

2021-08-20

星座对人性格的判断，就是巴纳姆效应呀～讲的笼统也似是而非

海林Lin

2021-08-20

引用：所以看上去是虱子使人健康，其实是体温高导致虱子不栖息在人身上。所以，岛上居民的结论应该是看到没有虱子的人应该让他去就医，因为他发生了疾病。在原始部落并没有体温计，这个结论的确可以帮助到他们，而不是盲目的相信“虱子让人健康”。

问题：这和相关因果倒置有什么区别呢？

展开